Response to Reviewer #3

We thank Reviewer #3 for her/his time and helpful suggestions to clarify methodological points in the manuscript. The comments made by the reviewer are in black, and our responses are in blue.

This paper presents a method to cope with the problem of using correlated signatures when calibrating rainfall-runoff model on regionalized signatures. This is a well-recognized problem often neglected in regionalization studies and the methodology proposed in this paper appears satisfactory to estimate the potential uncertainties stemming from using multiple (in)dependent signatures. The paper is well written, to the point and very convincing.

Below are some minor comments that, to my opinion should be addressed to improve the clarity of the text.

Figure 4 suggests that the performance of the model constrained by signatures is relatively poor (median NSEprob around 0.6) whereas synthetic flows are taken as the reference. I am a bit confused by this result. Does this mean that the selected signatures are not informative enough to constrain satisfactorily the model parameters or does this stem from the uncertainties brought by the regionalization of these signatures? What would have happened if the 'observed' signatures were used instead of the regionalized ones?

Authors' reply: The relatively 'poor' performance results from both reasons pointed by the reviewer, i.e. limited information has been provided to constrain the model parameters and also due to the uncertainties introduced by the regionalization procedure.

It is important to note that, for the synthetic case shown in Figure 4, the regionalized signature analogue was produced by imposing an error equal to the observation-likelihood function (page 5396, lines 24-25). For example, streamflow elasticity (SE) could only be poorly regionalized leading large errors in model predictions (we only achieved a coefficient of determination of about 0.20 for this signature, while for the others the coefficient of determination was usually above 0.75 - see Almeida, 2014, for more detail on this). The large errors of SE are consistent with the fact that the boxplot corresponding to SE used in isolation to condition the model (fourth boxplot from the top in Figure 4) presents the lowest median value when compared with the other four boxplots that use RR, BFI, SFDC and HPC to condition the model. This is an indication that the regionalization error has an important impact on the results. But, of course, this is not the only reason, given that different signatures bring different information, which can be more or less valuable depending on the performance measure that we are using to evaluate our results.

To address the last point raised by the reviewer, we re-calculated the performance measure NSEprob when 'observed' signatures are used instead of the regionalized ones. And we obtained NSEprob value distributions very similar to the ones shown in Figure 4 (see Figure R.1 below). This suggests that the uncertainty around the regionalized signatures value as well as signature information content are the key factors causing a 'relatively poor' performance. In the revised manuscript we have added the following text at the end of Section 3.3.1:

"It is worth noting that very similar results (not shown here) are obtained when instead of regionalized signatures (calculated by adding noise to the exact signature value), 'observed' signatures (the exact signature value) are used. This suggests that the uncertainty around the regionalized signatures values as well as signature information content are the key factors leading to the results shown in Fig. 4."

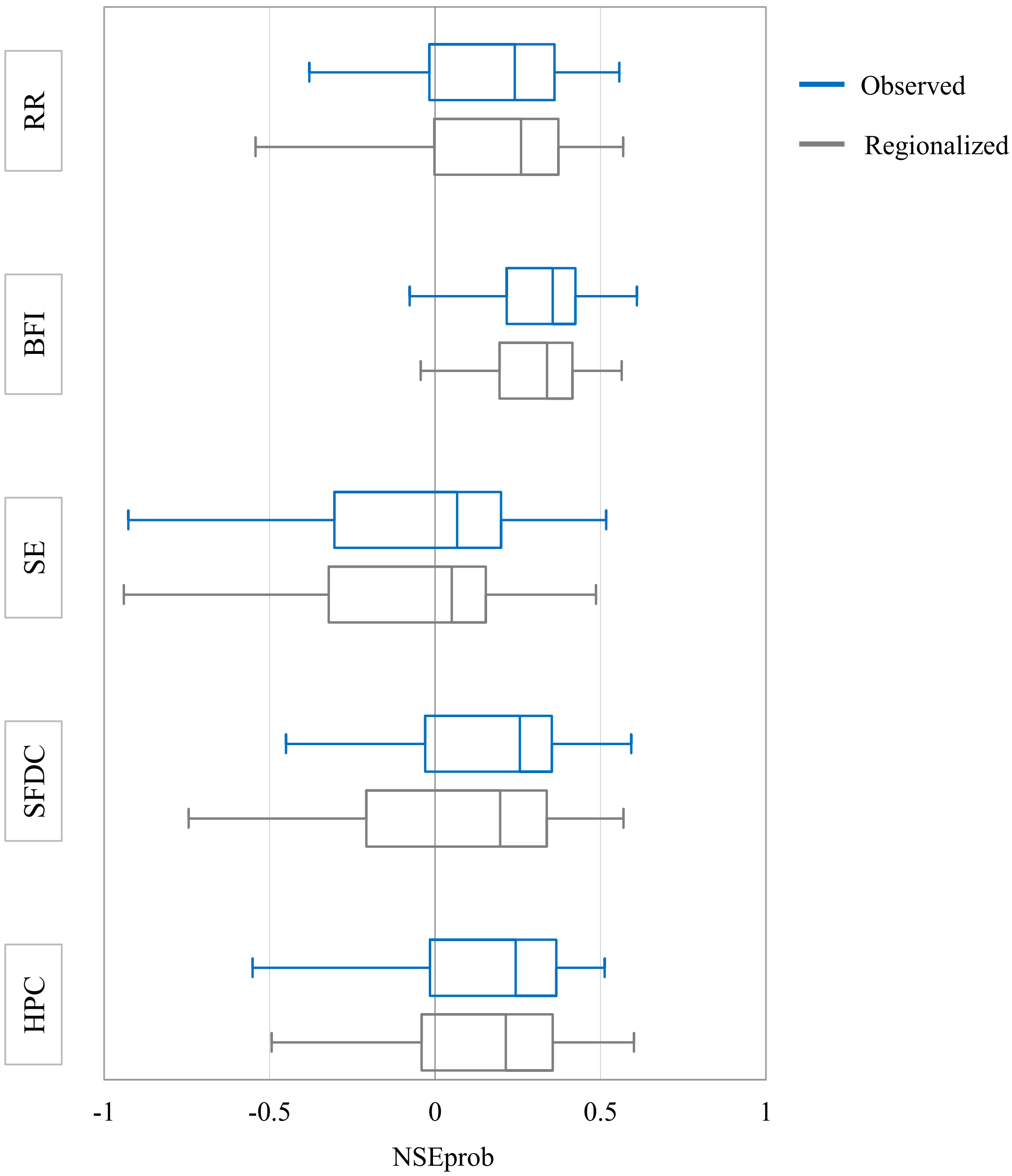
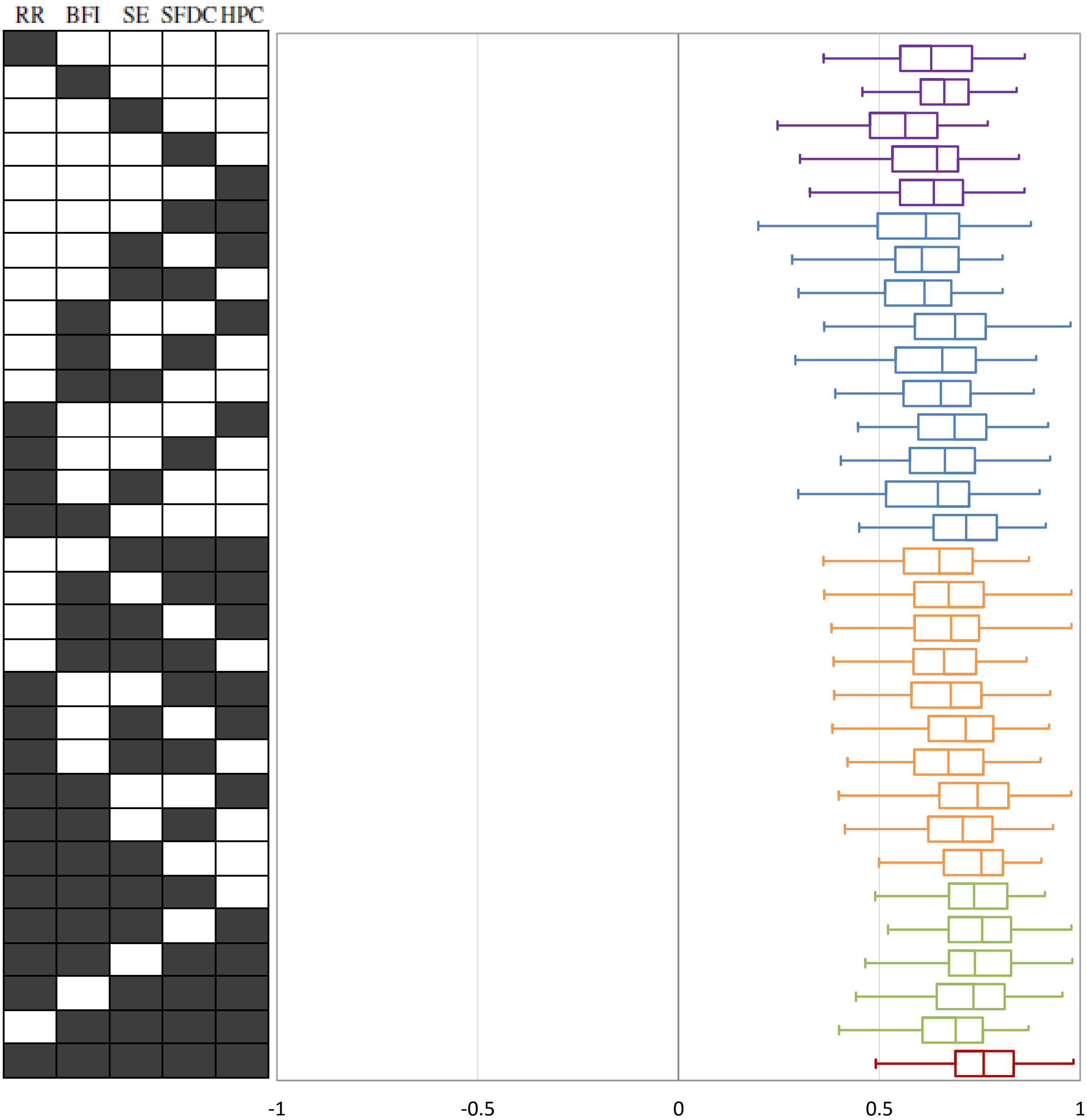


Figure R.1 - Boxplots representing the distribution of NSEprob values for each signature used in isolation for synthetic streamflow data. The blue boxplots correspond to the results obtained when 'observed' signatures were used (no noise added) and the grey boxplots correspond to the results obtained when the 'regionalized' signatures were used ('observed' signatures with added noise). The latter case is the same as presented in the original manuscript of our paper in Figure 4. The likelihood used is the same for both the blue and the grey boxplots and corresponds to the observed error structure derived as described in section 2.2.2 of the manuscript.

We would also like to stress that unlike the traditional NSE, the performance measure used in Figure 4 penalizes results for both the lack of accuracy and precision. So, if NSE measure is considered for the average prediction only, it will be higher than NSEprob of 0.6 (see Figure R.2 below).



$$NSE = 1 - \frac{\sum_{t=1}^{T}\left(E[\hat{q}_t] - q_t\right)^2}{\sum_{t=1}^{T}\left(q_t - E[q]\right)^2}$$

Figure R.2 - Boxplots representing the distribution of traditional NSE values for each combination of signatures for synthetic streamflow data. The colored boxplots correspond to the results obtained when inter-signature error correlations are considered in the likelihood function.

The description of the model should include at least the time step and the number of calibrated model parameters. Besides, it is not clear how the parameters are sampled from the posterior when generating an ensemble of flow simulations. Are the correlations between parameters taken into account in this procedure?

Authors' reply: In Section 2.3.1, page 5397, line 15, we state that 'daily time series (…) are employed.' However, we do not explicitly say how many parameters the model has or what the prior parameter ranges are. Instead, the reader is referred to Kollat et al. (2012) for further details on prior parameter ranges. For ease of reference, in the revised manuscript we have also provided a table that reproduces these model prior parameter ranges (see new Appendix A in the revised manuscript).

In terms of the sampling of parameters, 10000 parameter sets were sampled independently from a uniform distribution using the Latin Hypercube sampling so that probability of each parameter set is 10^{-4}. Then, to provide parameter samples that correspond to a uniform in signatures prior distribution, the parameter probabilities are re-weighted (see Almeida et al., 2013), and used in the further posterior distribution approximation. This allows accounting for correlation among the parameters due to the uniform in signatures prior. In the revised manuscript we have added the following text to the end of Section 2.2.1 to provide details about the parameter sampling techniques that were used:

"In other words, *N* parameter sets (*N* is equal to 10,000 in our study) are sampled independently from a uniform distribution using Latin Hypercube sampling, so that probability of each parameter set is 1/N (10^{-4} in our study). Subsequently, to provide parameter samples that correspond to a uniform in signatures prior distribution, the parameter probabilities are re-weighted (see Almeida et al., 2013), and used in the further posterior distribution approximation. This allows accounting for correlation among the parameters imposed by the uniform in signatures prior distribution."

The discussion proposed in section 3.4 is very interesting but could eventually be extended. With regards to the sensitivity of the results to the signatures used, I guess that the methodology presented in the paper does not allow removing uninformative signatures. One signature might be quite well regionalized but poorly informative for constraining the model and thus a methodology that gives more weight on well regionalized signatures might not be suitable in all cases. I fully understand that this is not the specific point discussed in the paper but since regionalization studies often focus on a specific flow range, the operational main question is which signatures are to be taken into account rather than how to avoid redundant information in the chosen signatures. . .

Authors' reply: This is a very interesting point. In fact this was one of the motivations to use streamflow elasticity, which could only be poorly regionalized. While for the other four signatures we obtained coefficients of determination that generally were greater than 0.75, for streamflow elasticity we only achieved a coefficient of determination of about 0.20 (for more detail, see Almeida, 2014). However, we do not analyze which signatures provide more information, so that possible redundant information can be discarded. Instead, in this paper we suggest that when we do not know which signatures are the most valuable (i.e. which ones are most informative), we should use them all while removing any duplicated information. This is particularly important, because well regionalized signatures may bring less information than poorly regionalized signatures, depending on what the focus of the study is. Discarding some of the poorly regionalized information on the grounds that it was poorly regionalized, therefore, may be a waste of resources. The method suggested in this paper enables all available regional information to be included, without omitting relevant information on a subjective basis or risking double-counting the same information.

References

Almeida, S. M. C. L.: The Value of Regionalised Information for Hydrological Modelling, Doctor of Philosophy, Department of Civil and Environmental Engineering, Imperial College London, London, UK, 2014.

Almeida, S., Bulygina, N., McIntyre, N., Wagener, T., and Buytaert, W.: Improving parameter priors for data-scarce estimation problems, Water Resour Res, 49, 6090-6095, doi:10.1002/wrcr.20437, 2013.

Kollat, J. B., Reed, P. M., and Wagener, T.: When are multiobjective calibration trade-offs in hydrologic models meaningful?, Water Resour Res, 48, doi:10.1029/2011wr011534, 2012.